Tree and Forestry Science and Biotechnology ©2010 Global Science Books

DATA*Musa* - a Database for Ortholog Genes from *Musa*

Roberto C. Togawa[1*] • Candice Mello Romero Santos[1] • Robert N. G. Miller[3,4] • Manoel Teixeira Souza Júnior[2] • Natália Florêncio Martins[1]

[1] Embrapa Recursos Genéticos e Biotecnologia. Parque Estação Biológica, CP 02372. Final W5 Norte, Brasília, DF, CEP 70.770-900, Brazil
[2] Embrapa LABEX Europe. Plant Research International – Wageningen University and Research Centre, P.O. Box 16, 6700AA, Wageningen, The Netherlands
[3] Universidade de Brasília, Campus Universitário Darcy Ribeiro, Instituto de Ciências Biológicas, Asa Norte, Brasília, DF, CEP 70.910-900, Brazil
[4] Universidade Católica de Brasília, SGAN 916, Módulo B, Brasília, DF, CEP 70.790-160, Brazil

Corresponding author: * **togawa@cenargen.embrapa.br**

ABSTRACT

Plantains and bananas (*Musa* spp.) are staple foods for rural and urban consumers in the tropics and an important source of income, particularly for smallholders. Given current threats to global *Musa* production caused by such biotic stresses, considerable input is being made in *Musa* breeding programs and genomics. The Global Musa Genomics Consortium (GMGC) has encouraged collaboration and genomics information and resource sharing between member institutes, ultimately for the development of improved cultivars for smallholder farmers. There are currently in the order of thirty thousand ESTs across members of the *Musa* genus in Genbank, 3000 nucleotides and 7000 are genome survey sequence records. The present work describes the Data*Musa* databank in which over nine thousand sequences from both genomes A and B have been stored and analyzed. The present work presents the database organization for all expressed sequence tag data. Sequences from different cDNA libraries were used to construct the database. Contig and singleton datasets were analysed via an in-house pipeline consisting of translation and alignment via Blast against several databases (non-redundant GenBank, Swissprot, KOG, GO and MIPs). Manual curation was used to annotate consensus sequences using an e-value cut-off of e^{-20} over a 40% coverage based upon data from at least three databases. The organization of DATA*Musa* enabled the identification of hundreds of new genes and provided information for further molecular studies, such as molecular marker development, promoter screening, gene expression profiling and molecular pathway analyses. All data is available at http://genoma.embrapa.br/musa/index.html, with access authorization granted by the DATA*Musa* management committee.

Keywords: data storage, gene annotation, stress responses, *Musa acuminata*, EST, assembly, transcribed genes

INTRODUCTION

Banana and plantains are grown mainly by smallholder farmers worldwide. Banana production provides an important source of income through national and international trade, and also represent an important staple commodity for key developing countries such as Brazil and India. According to estimations by the Food and Agriculture Organization of the United Nations (FAO), world total exports of banana accounted for 16.8 million tonnes in 2006.

The Global *Musa* Genomics Consortium (GMGC) is an international network of investigators committed to establishing *Musa* as a model crop for studies of comparative genomics and for gene discovery, leading to sequencing of the banana genome and the creation of new banana varieties. In terms of comparative genomics, *Musa* is seen as an ideal model for understanding genomic evolution in relation to biotic and abiotic stresses, in a polyploid, vegetatively propagated crop. The consortium currently brings together expertise from 38 institutions in 24 countries. Members are committed to close collaboration and agree to share materials and resources, including sequence data and enabling technologies. The aim of the GMGC is to elucidate the genome of *Musa* spp. to guarantee the sustainability of banana production and use as a staple food, through a better understanding the genetics and genomics of this genus. The information generated by the GMGC will provide improved strategies for genetic improvement. Brazil has participated in this initiative through several sequencing projects (http://genoma.embrapa.br/musa/en/index.html) directed towards expressed sequence tags (ESTs) for gene discovery (e.g. Santos *et al.* 2005), comparative genomics (Lescot *et al.* 2008), and resistance gene analog (RGA) characterization (Miller *et al*. 2008).

Musa is a member of the monocot order Zingiberales, a Commelinid lineage that diverged from the line leading to rice (Poales) in the mid-cretaceous period over 100 million years ago (Lescot *et al.* 2008). The *Musa* species *Musa acuminata* (AA genome) and *Musa balbisiana* (BB genome), both with 2n = 22 chromosomes, represent the two main progenitors of cultivated banana varieties. In the context of the estimated genome size of 600 Mbases, the GMGC databank currently holds only 4.053,526 bases from the A genome and 2.226,898 from the B genome, revealing that there is considerable sequencing work ahead.

There are currently (March 2009) in the order of thirty thousand ESTs across members of the *Musa* genus in Genbank, 3000 nucleotides and 7000 are genome survey sequence records.

The present work describes the Data*Musa* databank in which almost nine thousand sequences from both genomes A and B have been stored and analyzed. This databank offers a useful approach to overcome the current limited genomics information in *Musa*.

Construction and content

Data*Musa* is a data resource, sequence delivery and annotation system written in perl scripts as a web-based software tool. Three data types are currently contained within the database: ESTs (Santos *et al*. 2005; Miller *et al*. 2009), BAC sequences (Lescot *et al*. 2008) and NBS-LRR Resistance Gene Analogs (Miller *et al*. 2008). The EST data, which makes up the bulk of deposited sequence data, was generated through random clone sequencing from 10 distinct cDNA libraries constructed from diverse plant tissues

Received: 30 January, 2009. Accepted: 28 April, 2010.

Research Note

Table 1 cDNA Libraries information.

Species	Tissue	Stress	Reference
Musa acuminata var. Calcuta 4	Root *in vitro*	-	http://www.cenargen.embrapa.br/publica/trabalhos/bp109.pdf
Musa acuminata var. Calcuta 4	Green peel	-	http://www.cenargen.embrapa.br/publica/trabalhos/bp109.pdf
Musa acuminata var. Calcuta 4	Male flower	-	http://www.cenargen.embrapa.br/publica/trabalhos/bp109.pdf
Musa acuminata var. Calcuta 4	Leaves infected early state		http://www.cenargen.embrapa.br/publica/trabalhos/bp109.pdf
Musa acuminata var. Calcuta 4	Bulk	-	http://www.cenargen.embrapa.br/publica/trabalhos/bp109.pdf
Musa acuminata var. Calcuta 4	Leaves	Hot/cold	Santos *et al*. 2005
Musa balbisiana	Leaves	-	Santos *et al*. 2010*
Musa balbisiana	Root	-	Santos *et al*. 2010*

across *M. acuminata* and *M. balbisiana*, and during abiotic and biotic stress conditions (**Table 1**). Base calling and quality assignment of individual bases of the EST sequences were performed using the program PHRED (Ewing *et al*. 1998). All sequences with at least 100 nucleotides having a PHRED >= 20 were considered for clustering as described in Santos *et al.* (2005). In the current database release an approximate total of 8900 refined ESTs are deposited. The processing and annotation pipeline for data deposited in Data*Musa* takes sequences and quality output files from the sequencer, applies vector screening, quality trimming, clone tracking, and clustering to produce a set of unique sequences that are deposited named Musa assembled sequences (MAES) which are a combination of singletons and clusters of overlapping sequences.

1. DNA sequencing and processing

5' end DNA sequencing was conducted using T3 as sequencing primer. 5' sequencing allows assignment of functional annotation to as many transcripts as possible. Sequencing was performed, according to manufacturers' protocols, on both ABI 3700 and MegaBACE sequencers. Raw data was transmitted to the SisGen data management program (Pappas *et al*. 2008) and pre-processed. All pre-processed sequence fasta files were loaded into the database and submitted to the annotation pipeline. Tracefiles and chromatograms (raw data) were initially submitted to quality trimming, with 5' and 3' sequence vectors and poly-A (T) tails then masked. An initial selection of raw data was conducted based upon quality (PHRED value over 20) and length extension (over 150 bp). This process enabled separation of good quality insert sequences from rejected sequences, which were considered to fail on the basis of sequence quality, short inserts or long vector sequences. Quality trimming was performed during the submission using the Phred/Phrap package (Ewing *et al*. 2005). The SisGen system pipeline performed assembly with TGICL (Pertea *et al.* 2003). Fasta sequences were retrieved from the SisGen database using perl scripts, with each unigene receiving appropriate nomenclature based on the genome type, cDNA library, as well as additional appropriate features. The resulting contigs and singlet sequences were loaded into Data*Musa* and submitted to the annotation pipeline, which consisted of a sequence alignment using BLASTx against several databases. These comprised the non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples, which totaled 2,090,096 sequences, *Arabidopsis thaliana* proteins from MIPS (Mewes *et al*. 2008) with 25,458 sequences, the KOG database with 60,758 sequences (Tatusov *et al.* 2003), the Swissprot database with 158,240 sequences (Apweiler *et al*. 2004) and the Gene Ontology database (39,674 sequences) (Lomax 2005).

2. Gene annotation

Gene annotation was a consensus of the best hit in the BLAST searches (Altschul *et al*. 1997). The results were loaded into web-based tables. An automatic annotation was performed with the no-hits response. Where all the answers were no-hits, the gene annotation assigned was a 'NID' (or no identified genes). For each unigene sequence a webpage displayed the results of BLASTx against all the databases. Manual annotation assigned a gene name in the gene file based upon all information concerning the cDNA library, the assembly result, the contig/singlet name, BLASTx results with links to the original text result, a gene annotation name, a full length observation, an information field and specific annotator research.

Musa Assembled ESTs Sequences – MAESs (Santos *et al*. 2005), annotated as positive for orthologs of genes known to be related to biotic and abiotic stresses, where then checked for cDNA clones which potentially contained the entire coding region of the gene as well as having the longest 5′-end. Further analysis searching for the differential expressed sequenced was made. The Digital Differential Display was performed through SisGen (Pappas *et al*. 2008), which calculates the Audic-Claverie and Fisher statistics. The results are shown in **Table 2** of the supplementary material.

RESULTS AND DISCUSSION

Data*Musa* was developed using the PERL programming language. As EST sequences may contain a variety of contaminants, which must be removed before sequence annotation, the SisGen system of filtering, cleaning and masking was used for construction of the trimmed assembled dataset. Only long insert sequences with a quality assigned were selected and further grouped into clusters. Unigenes should typically represent the sequence of an expressed gene, with individual clusters treated as single genes. However, experimental data shows that some biological explanations may result in multiple contigs within a single cluster, as a result of paralogy or alternative splicing (e.g. Kim *et al.* 2008). To address such an issue, one should ideally compare the complete genome with EST data. However, in the case of banana, the whole genome is not yet available.

The web interface implemented allowed the semi-automated annotation of genes through the "annotation tables". The categorization allowed the selection of the most expressed candidate genes. Users wishing to further explore annotation details can do this via links to the source at http://genoma.embrapa.br/musa/index.html/. The user can also use the 'search' menu that holds two different query types; one is the key word on the BLAST result and annotation file, or the sequence name, contig name or sequence feature (**Fig. 1**). Users can also use the BLAST search to compare their own sequences with indexed sequences from Data*Musa*. The displays of search results allow a link to the original sequence data.

The categorization of the EST collection as shown in **Fig. 2** where it reveals a conserved core of largely essential eukaryotic genes related to physiological and cellular processes such as translation, ribosomal structure and biogenesis, and those related to post-translational modifications, protein turnover and chaperones. The bars indicate the percentage clusters assigned to each EST collection. Comparison of datasets reveals significant differences in the representation of genes within the various functional categories such as the genes related to energy production and conversion, translation, ribosomal structure and biogenesis and the category related to post-translational modification, protein turnover and chaperones as well as amino acid transport and

Table 2 Results from the Digital Differential Display with the Audic-Claverie and Fisher statistics.

Cluster name	MaaES	Blast best hit	Musa B	Musa A	Audic-Claverie	Fisher
CL7Contig8	MaAES_ALL1441	No Hit	74.0	2.0	8.26E-44	1.47E-41
CL5Contig2	MaAES_ALL1579	No Hit	158.0	153.0	3.22E-34	2.46E-31
CL4Contig1	MaAES_ALL0784	annexin	0.0	418.0	1.11E-14	6.31E-15
CL1Contig17	MaAES_ALL0361	metallothionein-like protein	0.0	294.0	8.12E-07	2.33E-05
CL42Contig1	MaAES_ALL1203	photosystem I reaction centre subunit psaN precursor	28.0	6.0	1.13E-02	7.10E-04
CL1Contig4	MaAES_ALL0504	Ribulose bisphosphate carboxylase small chain chloroplast precursor	16.0	482.0	1.58E-02	8.48E-03
CL5Contig4	MaAES_ALL0017	Glycine cleavage system H protein	22.0	4.0	1.38E+01	8.47E+00
CL31Contig1	MaAES_ALL4884	intracellular pathogenesis-related protein	20.0	10.0	5.94E+04	2.84E+05
CL2Contig4	MaAES_ALL0904	endoxyloglucan transferase	143.0	464.0	2.57E+08	1.20E+08
CL17Contig2	MaAES_ALL1439	pEARLI 1-like protein	18.0	17.0	1.71E+07	6.27E+08
CL171Contig1	MaAES_ALL0973	photosystem I reaction center subunit X psaK	8.0	1.0	3.69E+09	2.22E+09
CL1Contig9	MaAES_ALL0616	No Hit	56.0	560.0	4.66E+09	1.40E+11
CL152Contig1	MaAES_ALL0370	No Hit	8.0	3.0	6.06E+09	3.05E+11
CL148Contig1	Mbalbisiana_PKW_Root_018_F03_b	Cytochrome c oxidase	8.0	3.0	6.06E+09	3.05E+11
CL177Contig2	MaAES_ALL0343	No Hit	5.0	0.0	1.22E+10	8.10E+10
CL409Contig1	MaAES_ALL0916	lipid transfer protein precursor	5.0	0.0	1.22E+10	8.10E+10
CL186Contig1	MaAES_ALL0595	No Hit	6.0	1.0	1.29E+11	7.50E+10
CL33Contig1	MaAES_ALL0343	No Hit	16.0	22.0	2.09E+10	6.28E+10
CL12Contig1	MaAES_ALL0916	lipid transfer protein precursor	1.0	80.0	2.15E+10	5.83E+10
CL7Contig2	MaAES_ALL1459	No Hit	7.0	3.0	3.01E+10	1.47E+12
CL11Contig3	MaAES_ALL0058	germin-like protein	0.0	61.0	3.69E+11	8.93E+10
CL132Contig1	MaAES_ALL1406	immunophilin	8.0	5.0	4.12E+10	1.78E+12
CL358Contig1	MaAES_ALL3859	chlorophyll a/b-binding protein (cab-11)	5.0	1.0	7.45E+10	4.24E+11
CL366Contig1	MaAES_ALL3495	acyl carrier protein	5.0	1.0	7.45E+10	4.24E+11
CL8Contig1	MaAES_ALL0261	abscisic stress ripening protein	4.0	107.0	7.77E+10	1.56E+11
CL66Contig1	MaAES_ALL0464	No Hit	7.0	4.0	7.82E+10	3.50E+12
CL1Contig3	MaAES_ALL0098	ribulose 1,5-bisphosphate carboxylase/oxygenase	10.0	10.0	7.83E+10	2.78E+11
CL34Contig2	MaAES_ALL0062	glutaredoxin	13.0	18.0	1.06E+12	3.17E+11
CL17Contig3	MaAES_ALL0017	Glycine cleavage system H protein	3.0	6.0	6,077,102,535,342,750	1.45E+12
CL10Contig3	MaAES_ALL0034	light regulated protein	0.0	49.0	2.66E+11	4.86E+11
CL79Contig1	MaAES_ALL0984	photosystem I light-harvesting chlorophyll a/b-binding protein	9.0	10.0	2.74E+12	9.31E+11
CL7Contig3	MaAES_ALL1474	Bowman-Birk type proteinase inhibitor II	6.0	4.0	3.47E+11	0.001483926595015435
CL225Contig1	MaAES_ALL0748	60S ribosomal protein L11	4.0	1.0	4.22E+11	0.002339586578805745
CL6Contig4	MaAES_ALL0003	No Hit	0.0	44.0	6.07E+11	0.001159327841642632 5
CL1Contig14	MaAES_ALL0721	metallothionein-like protein	0.0	44.0	6.07E+11	0.001159327841642632 5
CL164Contig1	MaAES_ALL0333	No Hit	5.0	3.0	6.77E+11	0.003026301922063522 2
CL198Contig1	MaAES_ALL0318	No Hit	5.0	3.0	6.77E+11	0.003026301922063522 2
CL20Contig1	MaAES_ALL1558	ripening-associated protein	0.0	42.0	8.43E+11	0.001818821552282694 2

metabolism, in the category of nucleotide transport and metabolism genes and intracellular trafficking, secretion, and vesicular transport. Several clusters were grouped in the 'others' category with no indication of their corresponding function.

Digital Differential Display (DDD) was used to analyze the differential expression among the EST sequences from *Musa* genomes A and B. As a result it was found that at least 39 ESTs were identified as differential expressed, besides the size of the libraries. This effect was considered by the use of Audic and Claverie statistical analysis (1997). Among the identified genes mostly were from *Musa acuminata* calcuta 4 (Genome A), and annotated as annexin, metallothionein-like protein, proteins from the photosynthetic complex as well as lipid transfer protein, germin-like protein and immunophilin. Interestingly there were identified as differentially expressed the protein related to abscisic acid stress ripening protein. These results also revealed the paralogs for lipid transfer protein precursor and several unidentified candidates. The Audic and Claverie (1997) method is a rigorous significance test detects differentially expressed genes from relevant cDNA libraries and the DDD, as a useful and friendly tool, allowed the discovery of new candidate genes Therefore, these findings could be considered to gene expression validation studies through Q-PCR and plant transformation.

CONCLUSIONS

In this paper, we presented an organized EST collection from *Musa* transcriptome. Using different tissues, several cDNA libraries were produced and assembled to produce 1,677 contigs and 7,307 singlets where around 5,000 orthologs and paralogs were annotated (table 3). Annotations were based upon several different public databases and therefore provided greater assurance about gene function. Novel features of Musa transcriptome were associated with abiotic stress as well as amino acid transport and metabolism. The Data*Musa* databank is a banana-specific workbench for investigating ESTs in both A and B Musa genomes and is freely available to academic researchers registering on the website and obtaining approval for access

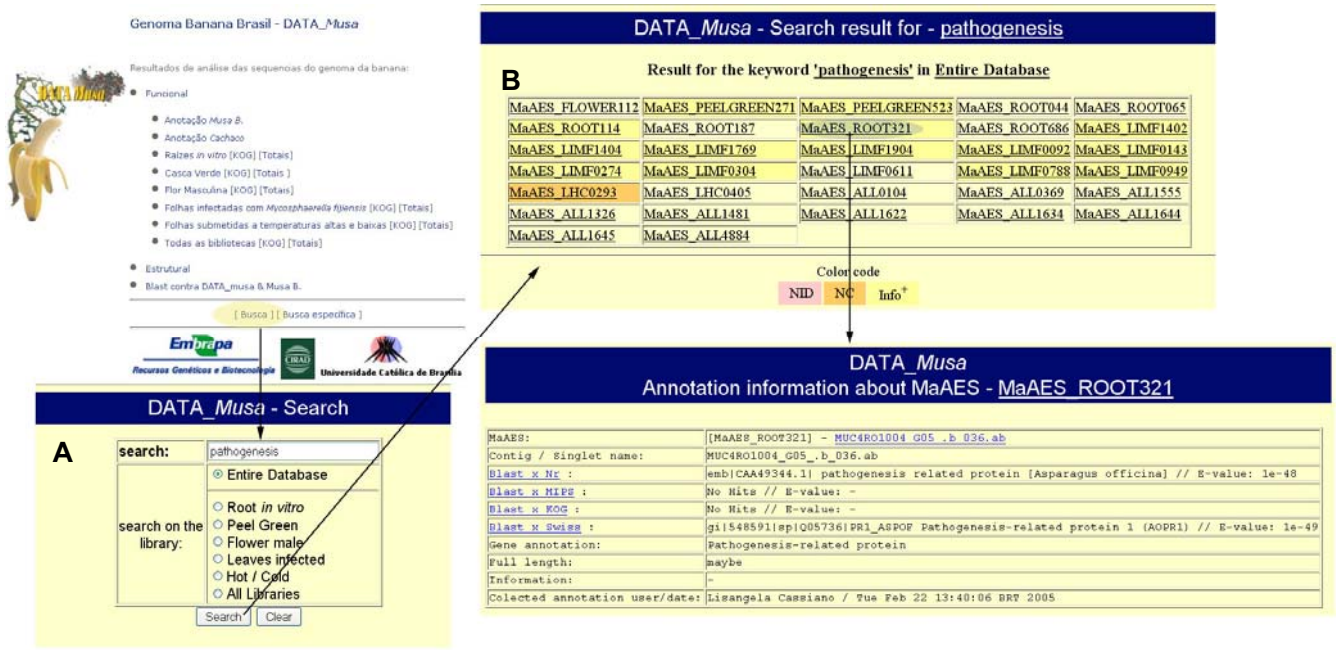

Fig. 1 Data*Musa* screenshots showing EST annotation data. Summary of EST annotations according to their functional classification, using a keyword search in the database.

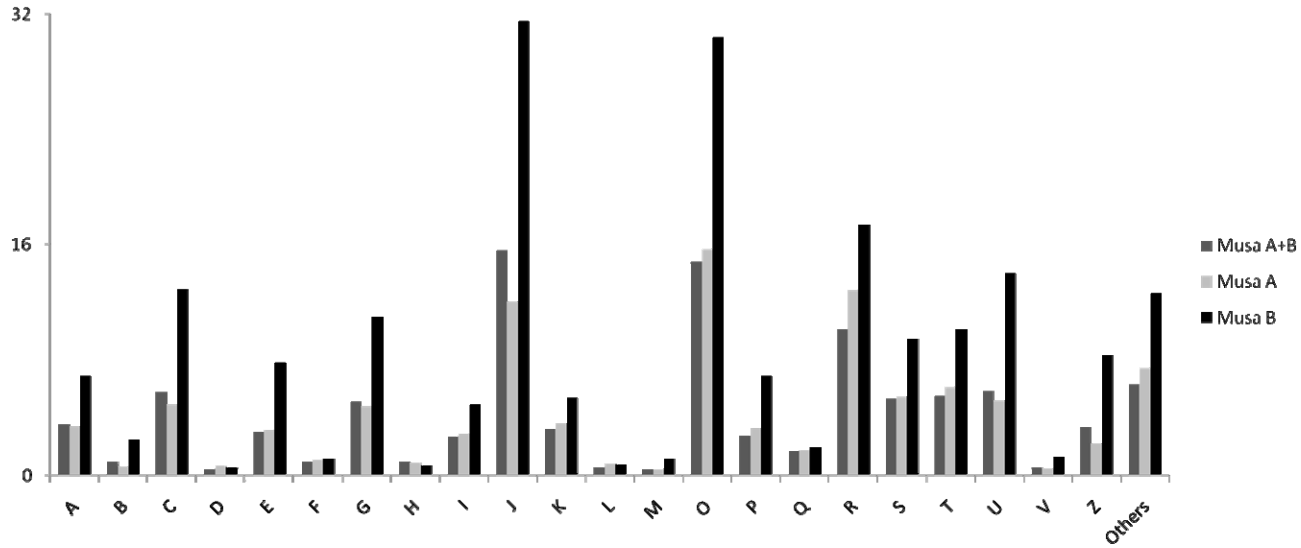

Fig. 2 Functional classification and *M. acuminata* assembled EST sequences, *Musa balbisiana* and *Musa* A + B assembling, according to eukaryotic clusters of orthologs (KOG). Designations of functional categories: M - Cell wall/membrane/envelope biogenesis, N - Cell motility, O – Post-translational modification, protein turnover, chaperones, T - Signal transduction mechanisms, U - Intracellular trafficking, secretion, and vesicular transport, V -Defense mechanisms, W - Extracellular structures, Y - Nuclear structure, Z - Cytoskeleton, A - RNA processing and modification, B - Chromatin structure and dynamics, J - Translation, ribosomal structure and biogenesis, K - Transcription, L - Replication, recombination and repair, A - RNA processing and modification, B - Chromatin structure and dynamics, J - Translation, ribosomal structure and biogenesis, K - Transcription, L - Replication, recombination and repair, C - Energy production and conversion, D - Cell cycle control, cell division, chromosome partitioning, E - Amino acid transport and metabolism, F - Nucleotide transport and metabolism, G - Carbohydrate transport and metabolism, H - Coenzyme transport and metabolism, I - Lipid transport and metabolism, P - Inorganic ion transport and metabolism, Q - Secondary metabolites biosynthesis, transport and catabolism, R - General function prediction only and S - Function unknown.

from the managing committee. Database content is also available through a sequence retrieve request at http://genoma.embrapa.br/musa/anotacao/.

ACKNOWLEDGEMENTS

This work was funded by the CNPq (Projects 680.398/01-5 and 506165/2004-3), the IAEA (Project 13187), FINEP (Project 0107060900 / 0842/07), Embrapa and the Universidade Católica de Brasília. C.M.R.S. was supported by a fellowship from the Fundação de Apoio a Pesquisa do Distrito Federal (FAP/DF). This work represents a part of the Brazilian participation in the Global *Musa* Genomics Consortium (http://www.musagenomics.org).

REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-3402

Apweiler R, Bairoch A, Wu CH (2004) Protein sequences databases. *Current Opinion in Chemical Biology* **8**, 76-80

Audic S, Claverie JM (1997) The significance of digital gene expression profiles. *Genome Research* **7**, 986-995

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research* **8**, 175-185

Kim HJ, Baek KH, Lee SW, Kim J, Lee BW, Cho HS, Kim WT, Choi D, Hur CG (2008) Pepper EST database: comprehensive *in silico* tool for analyzing the chili pepper (*Capsicum annuum*) transcriptome. *BMC Plant Biol-*

Table 3 cDNA libraries.

cDNA tissue	Contigs	Singlets	Total	Annotated	No hits
Musa acuminata Root *in vitro*	177	972	1149	612	437
Musa acuminata Green peel	167	780	947	595	352
Musa acuminata Male flower	140	838	978	560	418
Musa acuminata Leaves infected early state	507	2123	2630	1491	1139
Musa acuminata Leaves under heat/cold stress	217	802	1019	683	336
Musa acuminata Bulk	25	110	135	-	-
Musa balbisiana Leaves	201	765	966	751	215
Musa balbisiana root	243	917	1160	941	219
Total	1677	7307	8984	5633	3116

ogy **8**, 101-108

Lescot M, Ciampi AY, Ruiz M, Blanc G, Leebens-Mack J, Garsmeur O, D'Hont A, da Silva FR, Ronning CM, Cheung F, Haas BJ, Althoff R, Arbogast T, Hine E, Pappas Jr. GJ, Souza Jr. MT, Miller R, Glaszmann JC, Town CD, Piffanelli P (2008) Insights into the *Musa* genome: Syn-tenic relationships to rice and between *Musa* species. *BMC Genomics* **9**, 58-62

Lomax J (2005) Get ready to GO. A biologist's guide to the Gene Ontology. *Briefings in Bioinformatics* **6 (3)**, 298-304

Mewes, HW, Dietmann S, Frishman D, Gregory R, Mannhaupt G, Mayer K, Muensterkötter M, Ruepp A, Spannagl M, Stuempflen V, Rattei T (2008) MIPS: Analysis and annotation of genome information in 2007. *Nucleic Acids Research* **36 (Database issue)**, D196-201

Miller RN, Bertioli DJ, Baurens FC, Santos CM, Alves PC, Martins NF, Togawa RC, Souza Jr. MT, Pappas Jr. GJ (2008) Analysis of non-TIR NBS-LRR resistance gene analogs in *Musa acuminata* Colla: isolation, RFLP marker development, and physical mapping. *BMC Plant Biology* **8**, 15-23

Miller RN, Passos, MAN, Emediato FL, de Camargo Teixeira C, Ciampi AY, Togawa R, Martins NF, Amorim EP, Vilarinhos AD, Pappas Jr. GJ (2009) Candidate resistance gene discovery in *Musa-Mycosphaerella* host pathogen interactions. II Brazilian Symposium on Plant Molecular Genetics, 31/03/2009 – 03/04/2009, Buzios, Rio de Janeiro, Brazil

Pappas Jr. GJ, Miranda RP, Martins NF, Togawa RC, Costa MMC (2008) SisGen: A CORBA based data management program for DNA sequencing projects. *Lecture Notes in Computer Science* **5109**, 116-123

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **19 (5)**, 651-652

Santos CR, Martins NF, Horberg HM, Almeida ERP, Coelho MCF, Togawa R, Silva FR, Caetano AR, Miller RNG, Souza Jr. MT (2005) Analysis of expressed sequence tags from *Musa acuminata* ssp. *burmannicoides*, var. Calcutta 4 (AA) leaves submitted to temperature stresses. *Theoretical and Applied Genetics* **110**, 1517-1522

Santos CR, Ferreira CF, Cassiano LP, Coelho MCF, Togawa RC, Martins NF, Souza Jr. MT (2010) Transcriptome analysis of leaves and roots of *Musa balbisiana* var. Pisang Klutuk Wulung. In: Tripathi L (Ed) *Bananas, Plantains and Enset*. *Tree and Forestry Science and Biotechnology* **4 (Special Issue 1)**, 77-82

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41